The Genome Sequence of *Caenorhabditis remanei*: Enhancing the Utility of *C. elegans* as a Model Organism

John Spieth – Washington University School of Medicine
Sean Eddy – HHMI/Washington University School of Medicine
Paul Sternberg – HHMI/California Institute of Technology
Robert H. Waterston – University of Washington Medical School
Richard K. Wilson – Washington University School of Medicine

Rationale: The value of *Caenorhabditis elegans* as a model organism is unquestioned. The recent acquisition of whole genome shotgun sequence for *Caenorhabditis briggsae* (http://genome.wustl.edu/projects/cbriggsae/) proved extremely successful, both in terms of sequence quality and its impact on understanding the *C. elegans* genome. However, because of the evolutionary distance between *C. elegans* and *C. briggsae* (comparable to mouse/human divergence, though somewhat greater) and the extensive rearrangements, many important features remaining ill defined or unrecognized.

Here we propose sequencing the genome of *Caenorhabditis remanei*, a third nematode species whose lineage arose near the divergence point of *C. elegans* and C. *briggsae*. This sequence will allow triangulation for genome alignment, gene interpretation, promoter analysis, and identification of ncRNAs (non-coding RNAs). Marginally conserved sequences between *C. elegans* and *C. briggsae* can be given more attention if also conserved in *C. remanei*, other functional elements may emerge from statistical noise in a three-way comparison and additional orthology will be determined, all of which will enhance and expand the value of *C. elegans* as a model for understanding human health and disease, and basic biological processes.

Background: *C. elegans* has been a major model system for basic biological and biomedical research. It was the first animal for which a complete description of its anatomy, development and neural wiring diagram and exists. It was the first multi-cellular organism to have its genome sequenced (The *C. elegans* Sequencing Consortium, 1998), which has stimulated functional genomics in animal systems. It has provided a platform for the development of powerful downstream genomic resources such as genome-wide gene inactivation by RNAi (Kamath *et al*, 2003) and gene expression mapping (Kim *et al*, 2001).

The value of *C. elegans* as a model organism for understanding human health and disease has long been recognized (Ahringer, 1997). Over half of *C. elegans* genes have human orthologs, while ~42% of human disease genes have an homologs in *C. elegans* (Culetto and Sattelle, 2000). As of January 2003, there were 468 human disease genes that could be assigned orthologs in *C. elegans* based either on mutual best hit criteria or from the literature (Schwartz, Chan, personal communication). In spite of its relatively simple anatomy, many of the cell types associated with complex mammalian functions such as intestine, neurons, muscle and excretory cells can be recognized in *C. elegans*.

Furthermore, the worm has contributed to, and is often at the forefront, of our understanding of fundamental biological processes such as programmed cell death, signaling pathways, cell movement and polarity, sex determination and synaptic signaling to name a few.

The *C. elegans* community is comprised of over 2000 researchers in 434 laboratories around the globe. There are annual meetings (international meeting in odd years; regional meetings (East Coast, West Coast, Midwest, European and Asian, in even years). Over 1600 researchers attended the 2001 International *C. elegans* Meeting. Over 700 *C. elegans* papers were published in the last year (2002); this has been increasing steadily; in addition, many more papers use *C. elegans* sequence for comparative purposes. The community is well organized with a genome database, WormBase (http://wormbase.org/), which will store and display the data from this project (P.S. and J.S. are two of the four WormBase PIs); the *Caenorhabditis* Genetics Center (http://biosci.umn.edu/CGC/CGChomepage.htm), which freezes, stores and distributes strains (including *C. remanei*); the ORFeome project (http://worfcb.dfci.harvard.edu/), which has generated open reading frame clones for most *C. elegans* genes; the Gene Knockout Consortium (http://elegans.bcgcs.bc.ca/knockout.shtml), which generates deletion alleles of genes for user request.

Genome-wide comparisons between *C. elegans* and *C. briggsae* using the nearly 14 Mb of finished *C. briggsae* sequence has shown sequence conservation in protein coding exons, whereas introns and intergenic regions are largely divergent (Kent and Zahler, 2000). Conserved regions outside known genes in some cases correspond to known regulatory elements (Xue *et al*, 1992; Kennedy *et al*, 1993). Other regions, by virtue of their proximity to genes, are believed to be putative regulatory regions, but others are of unknown significance. The identification and experimental understanding of these conserved sequences in gene regulation, chromosome dynamics and other functions will lead to a better understanding of the function of specific sequence features in the human genome. The recently assembled whole genome shotgun sequence covering 98% of the *C. briggsae* genome will facilitate the discovery of important functional regions of the *C. elegans* genome, but for reasons discussed below this will be greatly enhanced with the ability to do three way comparisons.

Utility of the *C. briggsae* sequence: *C. briggsae* was selected as the initial genome for comparative analysis because of its manifest ability to reveal sequences in *C. elegans* experiencing purifying selection. *C. briggsae* is also a self-fertilizing hermaphrodite very similar in appearance to *C. elegans*. It has a genome size approximately the same as *C. elegans* distributed over the same number of chromosomes (six). Despite these superficial similarities, *C. briggsae* and *C. elegans* are more divergent at the sequence level than mouse and human.

The analysis of the *C. briggsae* genome is underway. WormBase has evidence for 19,522 *C. elegans* genes plus 1,776 alternately spliced variants. There are 22,912 *C. briggsae* predicted protein-coding genes and 1102 RNA genes (L. Stein, personal communication). Direct nucleotide comparison of the two genomes using the WABA program (Kent and

Zahler, 2000) produced a total of more than 1.5 M alignments spanning 65% of the *C. elegans* genome. By focusing on the 1:1 orthologs in this set almost 2000 blocks of co-linearity were revealed between the two genomes with a mean size of 23.3 kb (52 kb median). The longest collinear block was 812 kb (on the *C. elegans* X chromosome) and in all just over 50 Mb of the *C. elegans* genome was covered in these blocks. Using a reciprocal best match criterion combined with consideration of conserved co-linearity, a total of 11,007 ortholog gene pairs (58%) were identified. Many of the unassigned genes were members of large gene families, often with locally duplicated members and others were predictions unique to one or the other genome. More critically for genes or regions that fall outside of collinear segments or genes without established orthologs, meaningful comparisons are more difficult. Thus about 40% of the genome can only be approached through less powerful, more general comparisons.

The *C. briggsae* sequence is already being used to refine gene prediction and to discover new genes in *C .elegans*. Preliminary results using TWINSCAN (Korf *et al*, 2001) identified ~3000 potential new genes, of which 20% have recognizable Pfam (http://pfam.wustl.edu/) domains suggesting that >600 of these are truly new genes. Many of remaining 80% may as well be real genes, but clearly, conservation in a third species such as *C. remanei* would be a strong filter for these predictions.

Likewise, *C. briggsae* has proven useful in identifying alternative exons including previously unrecognized 5' exons. For example, D. Sherwood and P. Sternberg (unpublished) identified a nucleotide substitution associated with the *evl-5* locus that was not in an exon of the existing gene structure. A *briggsae-elegans* comparison narrowed 9 potential exons to three, one of which was confirmed by RNA analysis. A third species may have narrowed this to the true one. Comparison of *elegans* and *briggsae* supported a potential alternative 5' exon (and hence new promoter) in the *lin-3* EGF-like growth factor gene that was then verified by RNA analysis (Liu, 1999).

Another example of the utility of the *C. briggsae* sequence can be found in G-protein coupled receptor (GPCR) gene families, which are very poorly represented in EST collections. *C. briggsae* was not particularly useful in finding new GPCRs, that could usually be achieved with TBLASTN searches of the *C. elegans* sequence using already well-annotated GPCRs. Rather, the comparative aspect helped enormously when there was no close relative to the *C. elegans* gene and which also had a complicated gene structure. In such cases the conservation with *C. briggsae* was often the only way to identify exon/intron boundaries (H. Robertson, personal communication). Considerable species-specific evolution between *C. elegans* and *C. briggsae* is found in these families, such that orthologous relationships are difficult to determine with just two species. Robertson estimates that up to 25% for some GPCR families have undergone duplication since the divergence of *C. elegans* and *C. briggsae*.

One major use of *C. briggsae* sequence has been for the identification of regulatory sequences. A first step in defining important regulatory regions is a comparison of *C. elegans* and *C. briggsae* between orthologous genes. For example, in one study, Kirouac and Sternberg (2003) found regions of three genes sufficient to direct expression in

particular cells by a deletion analysis of the *C. elegans* genes. With the *C. briggsae* sequence now available, a retrospective analysis indicates that the *elegans-briggsae* comparison correctly identifies at the scale of ~100 nucleotides all the regions identified by deletion analysis, thus greatly accelerating the identification of functional regulatory regions. It should be emphasized that in these comparisons there is considerable noise, including false positives and false negatives. A more systematic attempt at identifying candidate regulatory elements in conserved, non-coding sequences is underway as a part of the more global analysis of the two genomes.

There are examples where *briggsae-elegans* comparisons have revealed regulatory elements in a UTR. Jan et al. (1997) compared the 3' UTR of *tra-2* between *C. elegans* and *C. briggsae* and identified an element that confers LAF-1-dependent translational regulation on the mRNA. Interestingly, this element is present in human *gli-1*, the homolog of *tra-2*.

There is a small *C. briggsae* community (largely a subset of the *C. elegans* community) that has started genetic and molecular genetic experiments with *C. briggsae*. There are over 500 mutant lines including those with defects in vulval development, dauer pathway and sex determination pathway (D. Baillie, P. Sternberg and colleagues, unpublished; Eric Haag, personal communication; J. H. Thomas, personal communication). Several of these mutants (e.g., *unc-4*, *lin-11*, *daf-4*) have been cloned by a candidate gene approach involving transformation rescue and sequencing, but others display novel phenotypes. A SNP map is being constructed based on the genome sequence, and thus will facilitate the positional cloning of novel loci. RNAi inactivation of other genes has allowed the probing of functional conservation and divergence (Rudel and Kimble, 2001; Ashcroft *et al* 1999; Stothard *et al* 2002).

The *C. briggsae* sequence has been a great aid in interpreting gene structure and in finding other conserved sequences of potential function in *C. elegans*, but this is clearly not applicable across the entire *C. elegans* genome and many important features remain ill defined with only two genomes to compare. Inference drawn from these analyses would be substantially strengthened by comparison with a third species. Boundaries of features could be sharpened and features with marginal significance would be either confirmed or rejected.

RNA gene finding by comparative analysis: There are several examples emerging where multiple comparisons of related genomes have been significantly more powerful than the comparison of just two genomes (Cliften *et al*, 2001; Cliften *et al*, submitted; Thomas *et al*, submitted) One particularly illustrative application of comparative genome sequence analysis is the identification of novel structural RNA genes.

The RNA gene finding program QRNA (Rivas and Eddy, 2001) works by statistically testing the pattern of mutation observed in conserved pairwise sequence alignments, probabilistically classifying conserved regions as "coding", "structural RNA", or "other". Coding conservation shows a distinctive pattern because of synonymous/non-

synonymous codon changes, and structural RNA conservation shows compensatory base-pair changes preserving a consensus secondary structure.

For QRNA to succeed, the two sequences need to be closely related enough that standard alignment programs (e.g. BLASTN) can recognize the homology and produce a reasonably accurate sequence alignment across a significant fraction of a coding exon or RNA gene; but distantly related enough that there is are sufficient mutations to distinguish the different patterns. The optimum for QRNA analysis is at about 85% sequence identity; above 90-95%, or below 75-80%, performance falls off rapidly. At this level of divergence many bases are still identical by descent in any pairwise comparison. But by combining multiple genomes, any one base becomes less likely to remain unchanged by chance.

In a study in which QRNA was used to identify several new structural RNA genes in *E. coli* (Rivas et al 2001), five different gamma proteobacterial genome sequences were used. The best single genome comparison gave only about half the results as the full comparison of five genomes. Many expressed *E. coli* RNAs were predicted in only a subset of the five comparisons, even for genomes from organisms at similar evolutionary distances. Essentially the same has been seen in another study in which QRNA has been applied to the yeast *S. cerevisiae* genome, using comparisons to five other *Saccharomyces* species (J. McCutcheon and S.R. Eddy, manuscript in preparation). In a search for human RNA genes with QRNA, mouse/human comparison alone seemed too noisy, but a three-way comparison of mouse, rat, and human genomes (requiring a homologous candidate RNA gene to be predicted in each of the three pairwise comparisons) appears to give sufficient statistical power to identify new RNAs (in a pilot study, 4 of 14 new candidate human small RNA genes were expressed in one or more tissues on Northerns; T.A. Jones and S.R. Eddy, unpublished).

In *C. elegans*, where only a single comparative genome at appropriate distance (*C. briggsae*) is currently available, QRNA performance has been disappointing. We have little confidence in the predictions from only a single comparison, and only 2 of 59 candidate loci tested so far by Northern analysis have been seen to be expressed as novel small RNAs (S.L. Stricklin and S.R. Eddy, unpublished). The situation appears similar to what we see in a single comparison of human/mouse, without rat. A second comparative genome sequence for *C. elegans* analysis should give us the ability to "triangulate" as we are doing successfully with human/mouse/rat.

RNA gene finding is of course not the sole reason to sequence a third nematode; it is just a concrete single example of one of the many ways in which comparative sequence analysis will inform the *C. elegans* genome. Other applications, such as regulatory site identification, have generally similar requirements for an "ideal" comparative genome: an organism with similar biology, gene expression, and gene regulation, not so diverged that insertion/deletion events have confused detailed alignments, with a divergence close enough to permit alignment of conserved regions, but far enough to accumulate sufficient single base substitutions to easily distinguish conserved from non-conserved nucleotide sequence. Any one pair of comparisons will be limited.

Rationale for *C. remanei*: What would be the ideal distance for such a third species? This depends in large part on which features are under study. The central challenges faced in exploiting the *C. elegans* sequence today are the complete and accurate identification of the regulatory elements and non-coding RNA genes**.**

Protein coding genes are actually reasonably well described today based on a combination of *ab initio* methods and experimental data, particularly EST and cDNA sequences (for about half the genes) and the refinement provided by the *C. briggsae* sequence in defining exon boundaries, in detecting alternative exons and even in overall gene structure.

The identification of other elements in the genome, in particular regulatory elements and non-coding RNA genes, is by contrast in its infancy. The *C. briggsae* sequence reveals many conserved features that are not protein coding (only about half the aligned sequences between the two genomes fall into coding exons). Current computational methods only partially help unravel their function.

Yeast probably provides the most extensive experience with comparative analysis in eukaryotes today, where the comparison of multiple sequences has been very successful in advancing the understanding of these cryptic elements. ncRNA genes can be detected by the correlated changes across secondary structures. Hundreds of regulatory elements were identified, of which 1/3 represent previously unknown motifs, and many new sites of previously known motifs were recognized. This may well represent the vast bulk of promoter motifs (Cliften *et al*, personal communication).

.

For the yeast analysis, the most informative species were about 35% divergent at neutral sites (4-fold degenerate third position codons). More distantly related species (60% divergent) were less useful because alignment across the intergenic regions was often unsuccessful. Promoter elements were often in different order and some had been lost while new ones appeared. Nonetheless, the distant species were of value in confirming the nature of predicted motifs from the analysis of the more closely related species (Cliften *et al*, personal communication).

Mammalian genome analysis is approaching the status of yeast. The mouse human comparison, with about 45-50% divergence at neutral sites, has been invaluable in discriminating true genes from false predictions and from pseudogenes in both species. The comparison has also been useful in refining gene structures, and new gene prediction algorithms have been designed to exploit the comparative information. Rat whole genome assemblies are now available and initial reports are that despite the fact that mouse and rat are more closely related to each other than to humans, the sequence is adding substantially to interpretation, especially for non-genic features.

More anecdotally, Thomas *et al* (submitted) have examined a limited number of regions across multiple mammalian species to investigate more thoroughly what might be learned by multi-species comparisons at varying evolutionary distances. Chicken has been useful

in identifying protein coding exons, but has not been useful for finding other more rapidly evolving or subtle features. Multiple mammalian species, with mouse-human representing the extreme, helped to saturate substitution differences at unselected sites, thereby highlighting conserved sequences, including promoter sequences.

The most important consideration in selecting the target organism is its sequence divergence relative to the two *Caenorhabditis* species for which genome sequences are already available, *C. elegans* and *C. briggsae*. An ideal choice will be at a distance comparable to that of other genomes used for successful comparative analysis of genomic features, such as the mouse and rat genomes for informing the human genome (Mouse Genome Sequencing Consortium, 2002; Baylor Genome Center, unpublished), or the genome sequences of *Saccharomyces sensu stricto* species (*S. bayanus*, *S. mikatae*, *S. kudriavzevii*) for informing *S. cerevisiae* (Cliften *et al* 2001).

Free living nematodes have been isolated and characterized by various investigators over the years (Sudhaus and Fitch 2002). Generally these studies have found three classes of worms with respect to *C. elegans*. 1) Other varieties of *C. elegans*, with less than 1% sequence variation from. 2) Other *Caenorhaditis* species with few changes in the 18S rDNA. 3) Other genera with substantially more changes in the 18S rDNA sequence.

The other varieties of *C. elegans* have been characterized in efforts to identify and exploit SNPs in genetic mapping. The so-called Hawaiian strain differs from the Bristol strain N2 by about 1 base per 1000. Interestingly, some other wild isolates appear to have a mosaic structure with parts of their genomes approaching the Hawaiian strain in difference rates and other parts more closely resembling Bristol. They are as a group too similar in sequence to be used in comparative analyses.

The other *Caenorhabditids* include *C. briggsae* and some male-female species, including *C. remanei*. *C. remanei*, *C. briggsae*, and *C. elegans* are nearly equidistant from each other. Phylogenetic trees built in a variety of datasets show a near-trifurcation. Most data favors a tree that has *C. remanei* slightly more related to *C. briggsae* than either is to *C. elegans*, as shown in the mitochondrial cytochrome oxidase II tree in Figure 1 (on the other hand, a male-female sexual system is considered ancestral in *Caenorhabditis*, and *C. remanei* is thus a gonochoristic (male-female) outgroup to the hermaphroditics *briggsae* and *elegans*). This divergence is roughly comparable to the human/mouse/rat genomes; the two rodent genomes are deeply diverged from each other and confer almost independent comparative evidence when used to inform the human genome. Other more divergent *Caenorhabditis* species have also been isolated, including strains PS1010 and CB5161. Ideally, we might prefer a species that was a little more closely related to *C. elegans* than *C. briggsae*, something more comparable to the mouse/human divergence, but no closer *Caenorhabditid* is currently known (other than independent isolates of *C. elegans* itself, such as the "Hawaiian strain" used for SNP mapping studies).

The other genera of nematodes are much more distantly related. As such, they are of interest in the evolution of nematodes and protein function (e.g., Sommer, 2001), but will be less valuable in identifying as yet unrecognized functional elements in *C. elegans*.

Intrageneric sequence diversity in known *Caenorhabditids* is fairly extreme, especially considering that these soil nematodes are barely distinguishable by morphology. Figure 2 shows a phylogenetic tree illustrating the relative sequence divergence among several selected nematodes, with a few other animals included for comparison. This includes other nematodes that are presently of more biological interest. Figure 2 also shows relative distance to *Pristionchus*, an important developmental model; *Meloidogyne*, a major plant parasite; or *Ascaris*, a medically important animal parasite (for a comprehensive nematode phylogeny see http;//www.nematodes.org; Blaxter *et al*, 1998). At present, the importance of informing a major model organism genome sequence outweighs other biological considerations.

SSU rRNA in the *briggsae/elegans/remanei* clade is slightly more divergent than the human/mouse/rat clade; whereas *Caenorhabditis* PS1010, for example, is almost as far from *C. elegans* as Xenopus is from human.

C. remanei appears to be the current best and logical choice. More fieldwork would likely find additional species, both closer to and more distant from *C. elegans*. However, given the amount of sampling that has already taken place (several isolates of *C. briggsae*, *C. remanei* and many of *C. elegans* have been found), it is unclear how readily new species would be recovered. Independent of the *C. remanei* sequencing proposed as an immediate goal by this white paper, we plan to isolate new *Caenorhabitis* species from the wild for possible future sequencing.

An example of how a third species, *C. remanei*, can complement the current *briggsae-elegans* comparison is found in an analysis of the *C. elegans* homeodomain gene *mec-3*. Xue et al (1992) identified potential cis-regulatory elements in the *C. elegans mec-3* promoter by comparison to *C. briggsae*, and went on to demonstate that the elements specifically bound both *unc-86* and *mec-3* proteins. They then showed that these elements are as well, if not better, conserved in *C. remanei* (Figure 3).

A preliminary analysis of *C. remanei* random reads showed that the number of reads falling in the rDNA repeat was roughly that expected from a genome size similar to *C. elegans* and *C. briggsae* with about 100 copies of the repeat. We have no reason to believe at this time that the size of the *C. remanei* genome is much different, but will make a more refined estimate during data collection and adjust the read number accordingly.

Choice of strain for sequencing: The Caenorhabditis Genetics Center (HTTP://www.cbs.umn.edu/CGC/) permanently stocks several different isolates of *C. remanei*, including the EM464, SB146, and VT733 isolates which have been most often used in the *C. elegans* community. Figure 1 shows that these three *C. remanei* isolates are closely related, and essentially at the same distance from *C. elegans*.

We believe the best strain for sequencing is SB146, the 1974 isolate (Sudhaus, 1974) of *C. remanei* ssp. *remanei*. There are nomenclature issues surrounding EM464 and VT733

(*C. vulgaris*, *C. vulgariensis*, *C. remanei ssp. vulgaris*) that make SB146 the cleanest choice.

To ensure homozygosity, we are inbreeding *C. remanei* (SB146) 16x, which is expected to be completed in spring 2003. The inbred strain will be permanently archived as a frozen culture at the Caenorhabditis Genetics Center.

Sequencing Strategy: The strategy that we propose to sequence the genome of *C. remanei* is similar the approach as we have used for the genome of *C. briggsae*.

The *C. briggsae* genome was sequenced using a combination of large insert clone mapping and whole genome shotgun (WGS) assembly. The map currently contains 188 contigs, made up of a mixture of fosmid (16,414) and BAC (17,855) clones. The WGS used some 2.3M sequencing reads, representing about 11 fold sequence coverage. Included in this set were 20,000 BAC ends to link the sequence assembly with the map. The assembly yielded 105.8 Mbp of sequence 5341 contigs with an N50 contig size statistic of 41kbps. Scaffolding of these contigs, using read pair information, results in 107.5 Mbp of scaffold length in 899 scaffolds with an N50 scaffold size of 474kbps. Using the position and orientation of the BAC end reads and the FPC map, these scaffolds were positioned onto the FPC map contigs, resulting in 142 ultracontigs spanning 102Mbp and 436 unplaced scaffolds containing 6Mbp (mostly highly repetitive). Comparison with finished sequences indicates the assembly covers 98% of the *C. briggsae* genome. An automated process is now being used to close many of the remaining gaps cheaply and efficiently.

For *C. remanei*, we would propose to generate 1.6 million sequence read-pairs from plasmid and fosmid clones to produce approximately 10-fold coverage of the genome. The ongoing improvements in WGS assembly and the almost exclusive use of the ***C. remanei*** sequence to inform the interpretation of the *C. elegans* sequence removes the need for a clone-based physical map. The relatively high coverage will yield almost as high a continuity as was achieved for *C. briggsae.* We will then use the automated directed approach mentioned above to close about two-thirds of the remaining gaps. Given the high degree of divergence of *C. remanei* from either of the other two genomes, relatively high continuity will be needed to align adequately many regions of the genomes. One improvement that we would make to the approach used for *C. briggsae* is to increase the proportion of fosmid-based sequence reads from 1% to 2% of the total. This increase should provide a better assembly framework with an increased overall long-range continuity of supercontigs, which again will be important for maximally aligning these genomes.

References:

Ahringer J. (1997). Turn to the Worm*!* *Curr. Opin. Genet. Dev* 7: 410-415.

Alexandersson M, Cawley S, Pachter L. (2003). SLAM- Cross-species gene finding and alignment with a generalized pair hidden Markov model, *Genome Research,* in press.

Ashcroft NR, Srayko MA, Kosinski ME, Mains PE, Golden A. (1999). RNA-mediated interference of a cdc25 homolog in *Caenorhabditis elegans* results in defects in the embryonic cortical membrane, meiosis, and mitosis. *Develop. Biol*. 206, 15-32.

Blaxter ML, De Ley P, Garey JR, Liu LX, Scheldeman P, Vierstraete A, Vanfleteren JR, Mackey LY, Dorris M, Frisse LM, Vida JT, Thomas WK. (1998). A molecular evolutionary framework for the phylum Nematoda. *Nature* 392: 71-76.

The *C. elegans* Sequencing Consortium, (1998). Genome Sequence of the Nematode *C. elegans*: A Platform for Investigating Biology. *Science* 282: 2012-2018.

Cliften PF, Hillier LW, Fulton L, Graves T, Miner T, Gish WR, Waterston RH, Johnston M. (2001). Surveying *Saccharomyces* Genomes to Identify Functional Elements by Comparative DNA sequence Analysis. Genome Research, 11; 1175-1186.

Culetto E, Sattelle DB. (2000). A Role for Caenorhabditis elegans in understanding the function and interactions of human disease genes. *Human Molecular Genetics* 9:869-877.

Jan E, Yoon JW. Walterhouse D. Iannaccone P, Goodwin EB (1997). Conservation of the *C. elegans tra-2* 3' UTR translational control. *EMBO J*. 16, 6301-6313.

Kamath RS, Fraser AG, Dong Y, Poulin G, Durbin R, Gotta M, Kanapin A, Le Bot N, Moreno S, Sohrman M, Welchman DP, Zipperfen P, Ahringer J., (2003). Systematic Functional Analysis of the *Caenorhabditis elegans* Genome Using RNAi. *Nature* 421: 231-237.

Kennedy BP, Aamodt EJ, Allen FL, Chung MA, Heschl MFP, McGhee JD. (1993). The gut esterase gene (ges-1) from the nematodes Caenorhabditis elegans and *Caenorhabditis briggsae. Journal of Molecular Biology* 229: 890-908.

Kent WJ, Zahler AM. (2000). Conservation, regulation. Synteny and introns in a large-scale *C. briggsae-C. elegans* genomic alignment. *Genome Research* 10: 1115-1125.

Kiroauc, M, Sternberg, P. W. (2003). cis-regulatory control of three cell-fate genes in vulval organogenesis of *C. elegans* and *C. briggsae*. *Developmental Biology*, in press.

Kim SK, Lund J, Kiraly M, Duke K, Jiang M, Stuart JM, Eizinger A, Wylie BN, Davidson GS. (2001). A Gene Expression Map for *Caenorhabditis elegans*. *Science* 293: 2087-2092.

Korf I, Flicek P, Duan D, Brent MR. (2001). Integrating Genomic Homology into Gene Structure Prediction. *Bioinformatics*, 17, S140-S148.

Liu J, Tzou P, Hill RJ, Sternberg PW. (1999). Structural requirements for the tissue-specific and tissue-general functions of the *C. elegans* epidermal growth factor LIN-3. *Genetics* 153: 1257-1269.

Moss EG, Lee RC, Ambros V. (1997). The cold shock domain protein LIN-28 controls developmental timing in *C. elegans* and is regulated by the *lin-4* RNA. *Cell* 88: 637-646.

Mouse Genome Sequencing Consortium (2002). Initial Sequence and Comparative Analysis of the Mouse Genome. *Nature* 420; 520-562.

Rivas E, Eddy, SR. (2001). Noncoding RNA gene detection using comparative sequence analysis. *Bioinformatics*, 2:8.

Rivas E, Klein RJ, Jones TA, Eddy SR. (2001). Computational identification of noncoding RNAs in *E. coli* by comparative genomics. *Current Biology*, 11:1369-1373.

Rudel D, Kimble JE. (2001). Conservation of *glp-1* regulation and function in nematodes. *Genetics* 157, 639-654.

Sommer RJ. (2001). As good as they get: cells in nematode vulva development and evolution. *Curr Opin Cell Biol*, 6:715-20.

Stothard P, Hansen D, Pilgrim D. (2002). Evolution of the PP2C family in *Caenorhabditis* : rapid divergence of the sex-determining protein FEM-2. *J. Molec. Evol*. 54, 267-282.

Sudhaus W. (1974). Zur Systematik, Verbreitung, Okologie und Biologie neuer und wenig bekannter Rhabditiden (Nematoda) 2. Teil. *Zool. Jb. Syst. Bd.* 101: 417-465.

Sudhaus W, Fitch D. (2002). Comparative studies on the phylogeny and systematics of the Rhabditidae (Nematoda). *Journal of Nematology* 33: 1-70.

Xue D, Finney M, Ruvkun G, Chalfie M. (1992). Regulation of the *mec-3* gene by the *C. elegans* homeoproteins UNC-86 and MEC-3. *EMBO J*. 11, 4969-4979.

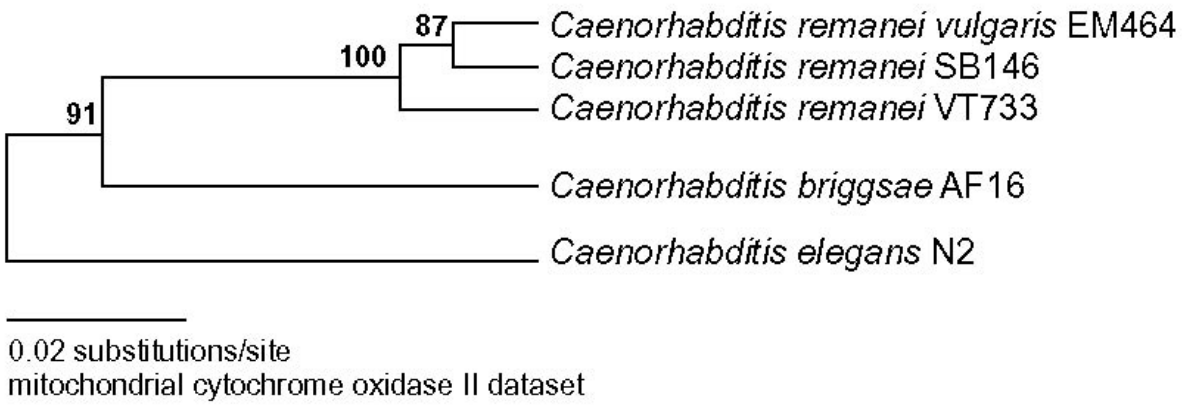

Figure 1:

Molecular phylogeny of *Caenorhabditis* species, inferred from an ungapped multiple alignment of a 687 nt fragment of mitochondrial cytochrome II genes [*C. remanei*: Genbank AF491511, AF491518,AF491519; *C. elegans*: Genbank NC_001328; *C. briggsae*: cb25 WGS assembly, (Washington University Genome Sequencing Center, unpublished), using Kimura two-parameter distances and the UPGMA algorithm, as implemented by PHYLIP 3.6a2 (J. Felsenstein). Support from 100 bootstrap replicates is shown in bold. UPGMA was used to root the tree; distances were sufficiently consistent with UPGMA's molecular clock assumption. Trees inferred from a *lin-28* coding sequence alignment (Moss *et al*, 1997) also slightly favor this (*C. briggsae, C. remanei*), *C. elegans* topology, while SSU rRNA trees instead slightly favor a (*C. briggsae, C. elegans*), *C. remanei* topology.

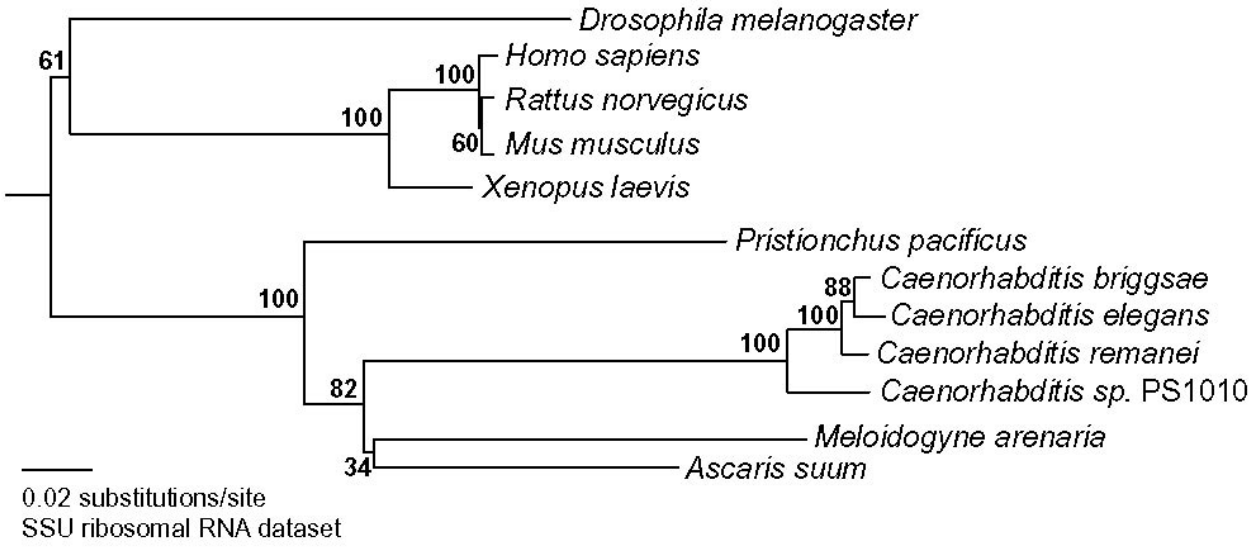

Figure 2:

Molecular phylogeny of selected nematodes and other metazoa. Tree was inferred from an SSU ribosomal RNA alignment of 2374 columns obtained from the Ribosomal Database Project (http://rdp.cme.msu.edu/html/), using Kimura two-parameter distances and the Saitou/Nei neighbor-joining algorithm (Phylip 3.6a2; J. Felsenstein), and rooted using *S. cerevisiae* (not shown). Support from 100 bootstrap replicates is shown in bold. What is relevant here are the relative distances between species; exact topological details of the relationship of the three mammals, the placement of *Drosophila melanogaster*, the placement of the deep branching nematodes, and the relationship of *C. briggsae*, *C. elegans*, and *C. remanei* are all insufficiently resolved by this dataset.

Figure 3: Conserved regions the promoter of *mec-3*.
Ce = *C. elegans*; Cb = *C. briggsae*; Cr = *C. remanei*. CS = conserved sequence. * = CDNA start.
From Xue *et al*, 1992.